ShadowSense: Unsupervised Domain Adaptation and Feature Fusion for Shadow-Agnostic Tree Crown Detection from RGB-Thermal Drone Imagery

Rudraksh Kapil Seyed Mojtaba Marvasti-Zadeh Nadir Erbilgin* Nilanjan Ray*
University of Alberta, Canada
{rkapil, seyedmoj, erbilgin, nray1}@ualberta.ca
*Equal Contribution

Abstract

Accurate detection of individual tree crowns from remote sensing data poses a significant challenge due to the dense nature of forest canopy and the presence of diverse environmental variations, e.g., overlapping canopies, occlusions, and varying lighting conditions. Additionally, the lack of data for training robust models adds another limitation in effectively studying complex forest conditions. This paper presents a novel method for detecting shadowed tree crowns and provides a challenging dataset comprising roughly 50k paired RGB-thermal images to facilitate future research for illumination-invariant detection. The proposed method (ShadowSense) is entirely self-supervised, leveraging domain adversarial training without source domain annotations for feature extraction and foreground feature alignment for feature pyramid networks to adapt domain-invariant representations by focusing on visible foreground regions, respectively. It then fuses complementary information of both modalities to effectively improve upon the predictions of an RGB-trained detector and boost the overall accuracy. Extensive experiments demonstrate the superiority of the proposed method over both the baseline RGB-trained detector and state-of-the-art techniques that rely on unsupervised domain adaptation or early image fusion. Our code and data are available: https://github.com/rudrakshkapil/ShadowSense.

1. Introduction

Forest environments are of great importance to ecosystems, economies, and society worldwide. A critical step in forest remote sensing is individual tree crown detection (ITCD), which can assist ecologists, foresters, biologists, and land managers in increasing the scope of their sampling for performing tasks such as forest health monitoring [5], pest infestation detection [20, 34], carbon storage estimation [7], and species identification [2, 39]. In recent years, various deep learning-based ITCD methods have been proposed to address the challenges in forest monitoring [63]. However, the lack of publicly available, diverse datasets tailored to this specific application has impeded progress in this research domain. Additionally, the ITCD task poses significant application-oriented and environmental challenges. These challenges encompass effectively harnessing the information from multiple sensors and ensuring the robustness of results in the presence of environmental factors. Existing tree crown detectors (e.g., [56]) have primarily been trained on RGB images, which are sensitive to occlusions and illumination variations (e.g., for shorter trees hidden in shadows). Nevertheless, the advantages of incorporating thermal images with complementary information in ITCD have been largely overlooked. While a few studies have used RGB-thermal data for urban tree crown detection (e.g., [37]), they require extensive manual pixel-wise annotation for supervised training and fail to address forest monitoring challenges, such as shadowed or occluded tree crowns. To bridge these gaps, this paper aims to provide an aligned RGB-thermal forest tree crown dataset and proposes a novel self-supervised approach that leverages both RGB and thermal imagery, improving the accuracy and adaptability of ITCD in various illumination conditions.

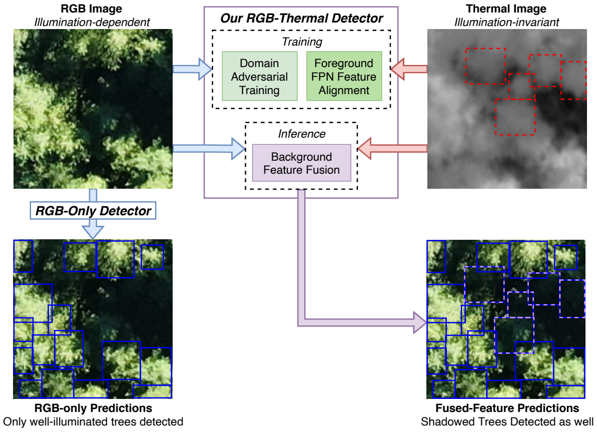

RGB Image
Illumination-dependent
Our RGB-Thermal Detector
Thermal Image
Illumination-invariant
Training
Domain Adversarial Training
Foreground FPN Feature Alignment
Inference
Background Feature Fusion
RGB-Only Detector
RGB-only Predictions
Only well-illuminated trees detected
Fused-Feature Predictions
Shadowed Trees Detected as well

Figure 1. **Overview of Proposed Method**. Undetected trees hidden in shadows are indicated by dotted red boxes. Best viewed in color.

The proposed method (ShadowSense) comprises domain adversarial training (DAT) and foreground (FG) feature alignment to learn domain-invariant representations and match observed tree crowns in both modalities (see Fig. 1). In particular, we train a shadow-agnostic ITCD model consisting of two parallel branches based on the RetinaNet architecture [28]. After initializing the branches with RGB-trained weights of the detector, we jointly train the thermal branch and three domain discriminators, minimizing the discrepancy over the feature extractor while maximizing it for domain discriminators. Tree crowns visible in both modalities are adapted by aligning FG feature maps of the feature pyramid network (FPN) using a simple yet effective intensity-based segmentation and morphological operations. During inference, the background (BG) regions of the FPN feature maps from RGB-thermal modalities are fused using a weighted average. The fused maps are then passed to the detector heads, leading to accurate prediction of tree crown bounding boxes. Our proposed method is entirely self-supervised, avoiding the need for labor-intensive manual annotations for model training. Moreover, we provide a challenging large-scale dataset consisting of undistorted and aligned RGB-thermal drone images, serving as a valuable resource to develop robust models and support future research.

The main contributions are summarized as follows.

(1) A novel shadow-agnostic tree crown detection method is proposed to exploit complementary information of RGB-thermal images and overcome the limitations of recent RGB-trained models for the ITCD task. This proposed method leverages the registered nature of available data for self-supervision (i.e., eliminating the need for data annotations) and incorporates source domain data post-adaptation.

(2) A challenging dataset for shadowed tree crown detection is provided to encompass varying degrees of shadows and illumination conditions of complex forest environments. This RGB-thermal dataset is large-scale and includes annotated images for evaluation and validation, as well as unlabeled images for training, aiming to advance the development of unsupervised/self-supervised methods.

(3) Extensive empirical evaluations validate the superior effectiveness of the proposed method over state-of-the-art (SOTA) methods utilizing image-to-image translation, early image fusion, or unsupervised domain adaptation (UDA).

2. Related Work

Tree Crown Detection. Deep learning methods have gained significant popularity in ITCD from RGB drone imagery in recent years. These methods primarily rely on well-known object detectors with different architectures [15, 63] and have found applications in various domains (e.g., [35]). However, these models are often trained on small datasets, resulting in moderate performance and their inability to effectively address challenges like overlapping tree crowns, small crowns, and distractors in various forest environments. Among the existing ITCD methods, DeepForest [56] stands out as the SOTA detector and was trained on a manually annotated dataset comprising over 10k tree crowns from 37 forests across the United States of America. Nevertheless, despite its remarkable performance in well-illuminated conditions, this RGB-only trained detector struggles to accurately detect trees with canopies hidden in shadows.

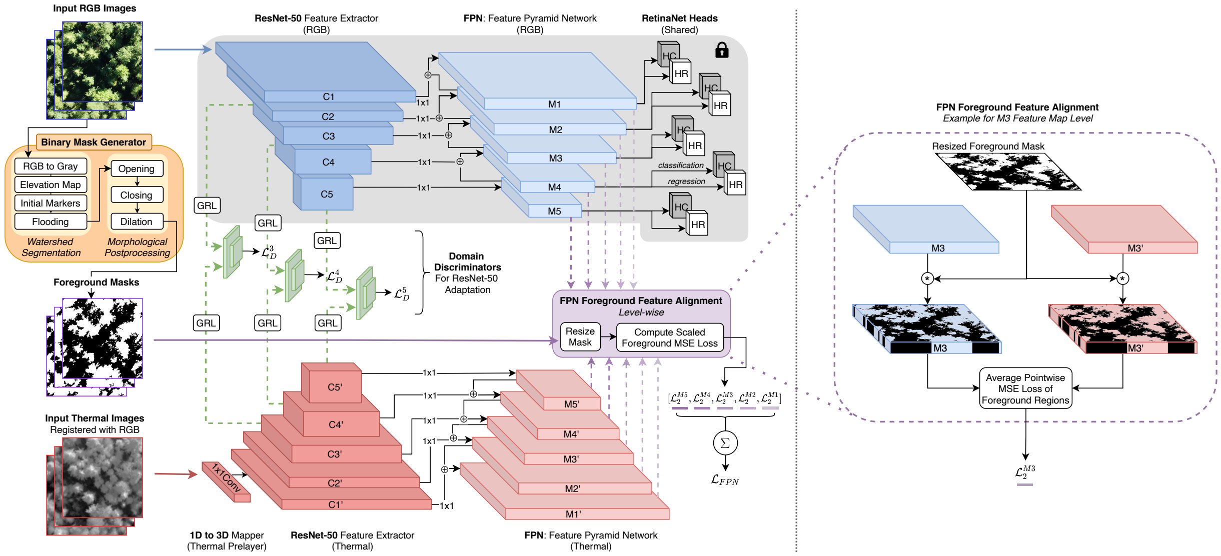
Unsupervised Domain Adaptation (UDA). The goal of UDA is to transfer knowledge from a source domain (e.g., RGB) to a target domain (e.g., thermal) without relying on annotations specific to the target domain [40]. In general, UDA involves adapting models either within the same modality (e.g., clear vs. foggy RGB) or across different modalities, such as RGB-thermal [1, 8, 23, 36, 42, 50, 65].

Many UDA approaches incorporate DAT by integrating domain discriminator networks into multiple parts of the model to encourage the learning of domain-invariant representations. These methods often employ global domain classification for the entire image [41] or local pixel-wise classification, focusing on FG regions [60] or other areas of interest predicted by attention modules [23, 50]. However, the potential application of these methods in the ITCD task is largely unexplored, despite their extensive development and success for generic object detection/segmentation. The drawback of existing UDA methods is their reliance on source domain annotations for training (see Table 1), which is not feasible in our problem setting. To address this limitation, the proposed method utilizes the registered nature of the available data for self-supervision during adaptation. Moreover, unlike previous approaches, our method retains and incorporates the source domain data after adaptation.

RGB-Thermal Early Fusion. Instead of adapting an RGB-trained model to the thermal data, an alternative approach is to fuse information from both modalities into a more informative image. Most works cope with the lack of ground truth (GT) fusion results by employing unsupervised RGB-thermal fusion methods. These methods can be applied to unregistered or registered images. For instance, UMFusion [52] improves upon existing methods for unregistered images by incorporating style transfer and a parallel-branch

Table 1. **Categorization of Related Works** according to training supervision through RGB ground truth (GT) annotations.

Category	GT Required	GT Not Required
UDA for Detection	[21] [22] [33] [36] [38] [41] [50] [51] [60] [65] [66]	**Ours**
RGB - Thermal Fusion	[10] [29] [31] [37] [64] SOD: [12] [27] [44] [48] [54] [61] [67]	[14] [30] [52] [55] [58] and **Ours**
Translation	[4] [11]	[17] [32]


Input RGB Images
Binary Mask Generator
RGB to Gray
Elevation Map
Initial Markers
Flooding
Watershed Segmentation
Opening
Closing
Dilation
Morphological Postprocessing
Foreground Masks
Input Thermal Images Registered with RGB
ResNet-50 Feature Extractor (RGB)
FPN: Feature Pyramid Network (RGB)
RetinaNet Heads (Shared)
C1
C2
C3
C4
C5
M1
M2
M3
M4
M5
HC
HR
classification
regression
1x1
GRL
$\mathcal{L}_D^3$
$\mathcal{L}_D^4$
$\mathcal{L}_D^5$
Domain Discriminators For ResNet-50 Adaptation
FPN Foreground Feature Alignment
Level-wise
Resize Mask
Compute Scaled Foreground MSE Loss
$[\mathcal{L}_2^{M5}, \mathcal{L}_2^{M4}, \mathcal{L}_2^{M3}, \mathcal{L}_2^{M2}, \mathcal{L}_2^{M1}]$
Σ
$\mathcal{L}_{FPN}$
1x1Conv
C5'
C4'
C3'
C2'
C1'
M5'
M4'
M3'
M2'
M1'
1D to 3D Mapper (Thermal Prelayer)
ResNet-50 Feature Extractor (Thermal)
FPN: Feature Pyramid Network (Thermal)
FPN Foreground Feature Alignment
Example for M3 Feature Map Level
Resized Foreground Mask
M3
M3'
Average Pointwise MSE Loss of Foreground Regions
$\mathcal{L}_2^{M3}$

(a) Complete Training Workflow

(b) FPN Foreground Feature Alignment

Figure 2. **(a) Detailed Workflow of Proposed Training Procedure** consisting of a thermal branch (in red) and an RGB branch (in blue). The weights of both are initialized from [56], and the latter's are frozen during training. The thermal feature extractor is trained to fool the domain discriminators (in green), and vice versa, using gradient reversal layers (GRL) at multiple levels. **(b) Close-up** of FPN feature alignment at the M3 level (in purple) that encourages foreground feature map regions in the two branches for a given image pair to match.

fusion module. For registered images, Wang *et al*. [55] propose an attention-based method for integrating thermal target perception and RGB detail characterization. These methods perform fusion at the image level, resulting in a new richer image with combined properties from both modalities.

Alternatively, fusion can be conducted at the intermediate feature level. Moradi *et al*. [37] propose a U-Net-based fusion model for tree crown segmentation, which requires GT segmentation maps for training. Supervised intermediate feature fusion methods have also been extensively studied in tasks like classification [25], segmentation [26, 45], and salient object detection [12,27,44,48,54,61,67] (see Table 1). In contrast, our method performs feature fusion during inference, rather than early fusion at the image level, in a self-supervised manner specifically designed for ITCD.

Image-to-Image Translation. Aside from fusion approaches, an alternative research direction involves colorizing a thermal image to resemble its RGB counterpart using encoder-decoder networks [4, 17, 32], and leveraging it for downstream tasks. Another approach is the use of classical algorithms [6] or SOTA deep learning methods [11] to translate RGB images into shadow-free versions. However, these methods typically discard the original RGB images in preference of the translated images, which may suffer from image artifacts and potentially contain less semantic information. Instead, our proposed method effectively fuses intermediate features extracted from both modalities after performing the UDA, thereby preserving the complementary information of RGB and thermal modalities.

3. Proposed Method and Dataset

In this section, "visible trees" refers to trees seen in both RGB and thermal images, primarily due to sufficient lighting conditions. In addition, "shadowed trees" are commonly shorter trees that remain hidden by the shadows of neighboring larger trees in the RGB image but become apparent in the thermal image. Due to the limitations of the illumination-dependent RGB modality, RGB-trained detectors are ineffective in identifying a significant number of shadowed trees. This is primarily because signals beyond the visible spectrum are undetectable using RGB alone. Hence, our proposed method first adapts the backbone of the existing baseline detector to the thermal data and then fuses extracted features from both modalities during inference. In the following, we present our proposed method and then introduce an RGB-thermal dataset that facilitates advancements in challenging shadowed tree crown detection and enables the development of robust models for the ITCD task.

3.1. Model Architecture and Training

Our network includes two parallel branches (i.e., RGB and thermal branches) based on the RetinaNet architecture [28] for the detection task (see Fig. 2a). Each branch comprises the backbone network, while the classification and regression heads that produce detection outputs are shared. The backbone includes a ResNet-50 network [16] that extracts features at multiple resolutions from input images, and an FPN that combines extracted features from multiple levels. We prepend a 1×1 convolutional layer (referred to as pre-layer) to expand thermal input images to three channels before passing them to the backbone network. We initialize the two branches with weights from a pre-trained RGB tree crown detector [56] and freeze the weights of the RGB branch to maintain its performance in the source domain. This is done because these weights effectively identify tree crowns from RGB images but perform poorly for thermal images, as indicated in Table 3. Then, we employ DAT and utilize FG FPN feature alignment to adapt the thermal branch to the target domain distribution. Considering the inherent low texture and contrast of thermal data, our training helps the thermal branch provide accurate predictions for visible and shadowed tree crowns.

Domain Adversarial Training (DAT). We employ DAT to train the feature extractor and thermal pre-layer to learn domain-invariant representations. Inspired by [41], we incorporate three domain discriminator networks (shown in green in Fig. 2) attached to the 3rd, 4th, and 5th levels of the extractor. These CNN classifiers predict the domain label (i.e., RGB or thermal) for the given feature map during training. Each discriminator is preceded by a gradient reversal layer (GRL) [9] that acts as the identity function in the forward pass, i.e., $G(\mathbf{x}) = \mathbf{x}$, but negates gradients in the backward pass. This layer ensures that gradients flowing through the extractor and discriminators are in opposition. This sets the stage for a two-player game: the feature extractor is trained to produce representations whose original domain is indistinguishable by the discriminators, while the discriminators aim to accurately classify the domain labels based on the feature representations.

We use the single-class focal loss to emphasize challenging images during DAT, as,

$$\mathcal{L}_D^c = -(1-p_t)^\gamma log(p_t), \tag{1}$$

where c $\in \{3,4,5\}$ is the level of the feature map, p_t is the predicted domain probability, and γ controls the diminishing rate of the modulating factor. A larger weight is assigned to more difficult instances, thereby increasing the importance of these challenging images in the overall loss calculation. Then, the game is modeled as a min-max optimization,

$$\min_{\{\theta_d^3,\theta_d^4,\theta_d^5\}} \max_{\{\theta_r,\theta_p\}} \mathcal{L}_D^3 + \mathcal{L}_D^4 + \mathcal{L}_D^5, \tag{2}$$

where $\theta_d^c, c \in \{3,4,5\}$ are the parameters of the three domain discriminators, θ_r are the parameters of the thermal feature extractor, and θ_p are the parameters of the thermal pre-layer. Unlike typical UDA works [38,41,50], we do not combine adversarial loss with a task-aware detection loss due to the lack of source annotations.

FPN Foreground Feature Alignment. It is crucial for FPN outputs of the RGB and thermal branches to align for the trees that are visible in both modalities (i.e., FG regions). This alignment acts as a proxy for task-specific loss to guide adaptation during training and is vital for the weighted average fusion process during inference (see Section 3.2). Thus, we can ensure the effective combination of complementary information from both modalities, leading to improved detection performance for shadowed tree crowns. Fig. 2b illustrates the alignment process for the third FPN feature map level. To do so, we first apply a binary BG/FG mask (described below) to the feature maps from the two branches, and then we compute standard average pixel-wise $L2$ loss between the residual values. Accordingly, five loss values denoted as $L_2^f, f \in \{1,2,3,4,5\}$ are obtained. These losses are then combined in a scaled manner, with higher weightage assigned to the larger feature maps using scaling values $\beta^f, f \in \{1,2,3,4,5\}$, i.e.,

$$\mathcal{L}_{FPN} = \sum_{f=1}^{5} \beta^f \mathcal{L}_2^f \ . \tag{3}$$

This alignment is complementary to the UDA process – both have the effect of producing the same feature maps at FG regions regardless of the modality. Therefore, $\mathcal{L}_{FPN}$ is used to update the parameters θ_f of the thermal FPN as well as the preceding parameters θ_r and θ_p.

To generate the binary masks used to train the detection model, we employ a simple yet computationally efficient method combining classic watershed segmentation [46] and mathematical morphology. This approach avoids the complexity of recent methods that utilize auxiliary neural networks for mask prediction. It is particularly suitable for our task because it leverages the assumption that BG pixels (including shadows) are generally darker than those in the FG. According to the binary mask generator in Fig. 2, we transform each RGB image to grayscale. Then, we mark all pixels with intensity $< \frac{20}{255}$ as 1 (representing the darker BG) and those with intensity $> \frac{100}{255}$ as 2 (i.e., brighter FG). We are confident about the FG/BG labels for these pixels, while those with intensities in between (initially unmarked) are determined through Meyer's iterative flooding algorithm [3], as implemented in scikit-image [43,49]. In this algorithm, an elevation map is computed using Sobel filtering. This map is then 'flooded' starting from the defined FG/BG markers. For this, each marked pixel's neighbors are inserted into a priority queue based on gradient magnitude, with enqueue time serving as a tiebreaker favoring the closer marker. The pixel with the highest priority is extracted, and if its already-marked neighbors share the same marker, it is assigned to that pixel. All unmarked neighbors that are not yet in the priority queue are enqueued. This flooding procedure iterates until the queue is empty and all pixels are marked as either FG or BG. After obtaining the initial binary mask, we apply three morphological operations for further refinement. we use 4-connected 3×3 structuring elements (SEs) to perform (1) *opening* to remove errant FG pixels surrounded by BG (2) *closing* to remove errant BG pixels surrounded by FG, and (3) *dilation* to pad FG boundaries and maintain FG performance during inference. Before using the binary mask, we resize it to match the dimensions at each FPN level.

3.2. Feature Fusion during Inference

During the inference phase, we exploit complementary information from the thermal branch to address the limitation of detecting shadowed tree crowns and improve the overall ITCD performance. This information resides in the BG regions of the RGB modality, which are typically prominent in the thermal modality. To achieve this, we follow the binary mask generation process used in the training phase, but now we assign "1"s to denote BG regions and "0"s for FG. Subsequently, we fuse the feature maps extracted from the RGB and thermal modalities in a level-wise manner. Fig. 3 illustrates this fusion process for the M2 level of feature maps.

While the FG pixels (depicted as black regions in Fig. 3) from the RGB feature maps are directly utilized, we mask the BG regions of the thermal feature maps to focus solely on the areas that are not visible in the RGB modality. As a result, the fused feature map F^f_{Fused} at level f is obtained through a weighted average of the RGB feature map (F_{RGB}) and the thermal feature map (F_T) for all BG pixels (x, y) as,

$$F^f_{Fused}(x,y) = \frac{F_{RGB}(x,y) + (F_T(x,y) \times \lambda_T \times \eta^f)}{1 + (\lambda_T \times \eta^f)}, \tag{4}$$

where λ_T is the weight assigned to thermal features for all levels and η^f denotes the fusion weight scaling specific to that level. η^f decreases with f because larger feature maps have a higher spatial resolution, and thus the averaging is less error-prone due to containing more fine-grained information. Once the fused feature map is obtained at each level, it replaces the RGB feature map and is fed into the classification and regression heads to predict bounding boxes.

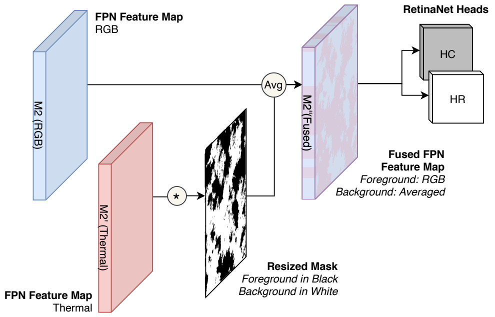


Figure 3. **Masked Fusion During Inference**, for the M2 level feature maps as an example. Background features (purple) are obtained by weighted averaging of the RGB (blue) and thermal (red) features. Foreground features are assigned the original RGB values. Best viewed in color.

Table 2. **Comparative Overview** of RGB-Thermal Image Datasets.

Dataset	# Pairs	Dimensions	Year	GT	Application
TNO [47]	63	Various	2014	×	Image Fusion
MFNet [13]	1569	640×480	2017	✓	Semantic Segmentation
VIFB [62]	21	Various	2020	×	Image Fusion
RoadScene [59]	221	768×576	2020	×	Image Fusion
LLVIP [18]	15488	1080×720	2021	✓	Pedestrian Detection
M^{3}FD [29]	4200	1024×768	2022	✓	Object Detection
RT-Trees (Ours)	49879	500×500	2023	✓(eval.)	Tree Crown Detection

3.3. Dataset for Shadowed Tree Crown Detection

In this section, we present an RGB-thermal dataset titled *RT-Trees* for advancing shadowed tree crown detection and developing robust models for ITCD. We built on an existing dataset [19] by conducting additional flights using the same setup. Specifically, we employed a DJI H20T sensor to capture RGB-thermal drone imagery during nine flights over a mixed forested region of central Canada. This data was then combined with available data from the five flights detailed in [19]. During data collection, we purposely diversified flight times to encompass a spectrum of challenging illumination conditions. Additionally, ever-changing climatic conditions throughout the year (e.g., temperature and snow cover) introduce an additional layer of diversity and challenges, especially in the more sensitive thermal images.

We proceeded with a series of preprocessing steps on the raw drone imagery, including cropping, resizing, co-registration, splitting into training/validation sets based on GPS coordinates, and providing high-quality annotations for evaluation. This resulted in a substantial collection of approximately 50k registered image pairs across all flights, signifying a considerable expansion compared to existing RGB-thermal datasets (see Table 2). We sampled and annotated 63 non-overlapping images for testing and 10 for validation from a single flight date. Each tree crown only appears once in these sets to ensure the reliability of performance evaluation, and the annotations differentiate between visible and shadowed (i.e., "difficult") tree crowns. The remaining bulk of images (49,806) was designated for training. These images display a high degree of overlap ($> 75\%$) and span all flights, a deliberate choice aimed at promoting diversity and consequently justifying the discrepancy in data split numbers. RT-Trees is primarily intended for self-supervised RGB-thermal ITCD, so no training set annotations are provided, but our proposed method demonstrates that the co-registered images can facilitate feature fusion methods. A notable characteristic of the RT-Trees dataset is the highly dense spatial distribution of detection targets compared to existing datasets, averaging around 60 tree crowns per image. Moreover, the presence of different tree species results in considerable variability of crown areas and shapes. We substantiate the challenges of RT-Trees with descriptive statistics and detail the collection, preprocessing, and annotation procedures in the supplementary material.

4. Experiments

In this section, we first provide implementation details for the proposed method (ShadowSense). We then compare its performance with the baseline and existing SOTA methods through the quantitative results reported in Table 3. We utilize three metrics for evaluation: (1) AP50, representing the average precision at 50% IoU (Intersection over Union) threshold, (2) AR100, representing the average recall over several IoUs given 100 detections, and (3) Percentage of correctly identified shadowed trees. The third metric focuses only on the difficult boxes, counting a positive if a prediction with an overlap of 85% with the BG regions was assigned to a difficult box. Finally, we present qualitative comparisons to support our experimental findings.

4.1. Implementation Details

We employed the well-known RGB-trained crown detector DeepForest [56] as our baseline method to demonstrate the effectiveness of the proposed method. The RetinaNet networks [28] in each RGB/thermal branch were initialized with pre-trained weights from [56]. To ensure fair comparisons, we configured the RetinaNet hyperparameters similarly to those employed in our baseline. This involved setting the non-maximum suppression threshold to 0.15 and the score threshold to 0.1 (default in [56]). During training, we set FPN alignment scales $\beta = [1.0, 1.0, 0.5, 0.05, 0.01]$ and the focal loss parameter γ to 2 (recommended in [28]). The domain discriminators consisted of three *Conv-BatchNorm-ReLU-Dropout* layers, an adaptive average pooling layer to reduce feature maps to a single channel, and a linear layer to finally produce a single output representing the confidence of belonging to the target domain. Dropout layers with a probability of 0.5 were included for regularization. To suppress noisy classification signals during early training stages, we gradually increased the GRL adaptation factor from 0 to 1, as prescribed in [9]. A training batch size of 16 was used in all experiments. The Adam optimizer [24] was used with an initial learning rate of 0.001 that was exponentially decayed with a gamma factor of 0.9 after each epoch (i.e., a complete pass through the training set). The training was conducted for 10,000 iterations, a sufficient period to observe plateauing in all training losses. The implementations were performed on a single Nvidia GeForce RTX 3090 GPU with 24-GB RAM. During inference, we performed weighted fusion using a thermal weight of $\lambda_T = 5$, which provided the best results. Similar to β, the scaling weights $\eta = [1.0, 1.0, 0.5, 0.2, 0.2]$ were applied to weight more towards thermal features in larger feature maps, while also ensuring that all products of λ_T and η are greater than or equal to one (i.e., always at least equal weighting between thermal and RGB features). Further validation of selected hyperparameters is provided in the supplementary material.

4.2. Baseline Quantitative Comparison

We evaluated the performance of the baseline model [56] in four scenarios. The first two involved assessing the effectiveness of the off-the-shelf model on RGB images and thermal images, respectively. The performance on RGB images was 49.86% AP50 and 24.01% AR100, although only 10.41% of difficult shadowed trees were successfully identified. When using thermal images, the baseline detector exhibited significantly inferior performance (see Table 1). The results demonstrate that the off-the-shelf RGB-trained detector is ill-suited for the thermal domain. In the other two scenarios, we conducted supervised fine-tuning of the detector model on RGB imagery, using supervised focal loss [28] for 10 epochs, following [56]. We used a subset of RT-Trees comprising 326 non-overlapping images containing over 22.5k crowns of visible and shadowed trees, which we manually annotated by inspecting the RGB-thermal pairs. The performance on RGB images shows a lead of 5.34% and 5.41% in terms of AP50 and AP100, respectively, while also resulting in a 9.69% increase in the detection of shadowed trees. Although the thermal modality is not directly used for training, this configuration requires costly annotation based on both modalities. Also, the performance of this model on thermal images is dramatically poor due to low spatial resolution and lack of fine details in these images. Instead, our ShadowSense can achieve superior performance by leveraging multi-modal data without needing *any* annotations.

4.3. State-of-the-art Quantitative Comparison

We compare the performance of ShadowSense with various SOTA methods that utilize image-to-image translation, RGB-thermal early fusion, or UDA. The baseline detector was applied to the generated images in the image translation and fusion methods, and we adopted the proposed weighted-average fusion of BG FPN feature map regions in all UDA experiments (including ours) to ensure fair comparisons. Additionally, an ablation study was conducted to analyze the impact of different components on our method.

Image-to-Image Translation. We investigated three methods: PearlGAN [32] (SOTA thermal colorization); ShadowFormer [11] (SOTA shadow removal); and a classic method that increases the brightness of pixels in HSV color space proportionally to their original brightness, i.e., darker pixels are made brighter. Thermal images colorized using PearlGAN performed worse than the baseline by -9.14% AP50 and -4.34% AR100. However, slightly more shadowed trees were detected. The decreased performance can be attributed to the introduction of artifacts and an overall loss of semantic information compared to the original RGB images. The detection performance on images generated by ShadowFormer was the most inadequate. The classical shadow removal method showed better results than PearlGAN but still performed worse than the baseline. This method jitters

Table 3. **Quantitative Comparison** of the proposed method with baseline and SOTA methods based on AP50 and AP100 metrics. The best and second-best results are emboldened in red (supervised) and blue (self-supervised).

Evaluation		Method	Training Data	All Trees		Shadowed Trees
				% AP50 (↑)	% AR100 (↑)	% Identified (↑)
Baseline [56]		Off-The-Shelf Model (Eval. on RGB Images)	Pre-trained on NEON [57]	49.86	24.01	10.41
		Off-The-Shelf Model (Eval. on Thermal Images)		4.34	2.27	2.82
		Supervised Fine-Tuned Model (Eval. on RGB Images)	+ Ann. RGB subset of RT-Trees	**55.20**	**29.42**	**20.10**
		Supervised Fine-Tuned Model (Eval. on Thermal Images)		5.64	3.98	3.81
Image Translation Inference using [56] on Generated Images		Increased Background Brightness in HSV Space	N/A	42.01	20.35	10.41
		ShadowFormer [11]: Shadow Removal	ISTD [53]	11.62	6.48	3.47
		PearlGAN [32]: Thermal Image Colorization	RT-Trees	40.72	19.67	11.18
Early Image Fusion Inference using [56] on Generated Images		UMFusion [52]	TNO [47]	38.00	18.56	10.20
		MFEIF [30]	TNO [47]	39.62	18.95	15.40
		MetaFusion [64]	M^3FD [29]	43.17	21.29	18.00
RGB-Thermal Unsupervised Domain Adaptation (UDA) w/o Source Annotations Our Fused Inference after Adaptation		SSTN [38]: Contrastive Learning	RT-Trees	31.53	15.16	2.13
		Attention-based UDA [50]	RT-Trees	31.97	15.34	3.84
		DA-RetinaNet [41]: ResNet DAT	RT-Trees	32.88	15.72	3.65
	Ablation Study	(i) **Ours**: FG FPN FA	RT-Trees	47.11	22.48	5.21
		(ii) **Ours**: ResNet DAT + FPN FA w/o Masking	RT-Trees	49.75	23.22	10.63
		(iii) **Ours**: ResNet DAT + FG FPN FA (Pred. Masks)	RT-Trees	52.18	24.84	9.33
		(iv) **Ours**: ResNet FG DAT + FG FPN FA	RT-Trees	52.24	24.38	14.32
		(v) **Ours** (**ShadowSense**): ResNet DAT + FG FPN FA	RT-Trees	**54.13**	**25.76**	**19.09**

the entire image inconsistently with the detector, leading to poor performance. Translation methods are thus ineffective for removing the shadows of dense canopies in our dataset.

RGB-Thermal Early Fusion. We evaluated three SOTA methods: UMFusion [52], MFEIF [30], and supervised MetaFusion [63]. MetaFusion directly generates a fused RGB image, while UMFusion and MFEIF convert the RGB image to the YCbCr color space, fuse the brightness (Y) channel with the thermal image, and then convert the fused image back to RGB. In all methods, shadowed trees become partially visible in the fused images to varying extents. According to Table 3, MetaFusion performed the best. Although the AP50 and AR100 results were still lower than those of the baseline detector, the fused images revealed 7.59% more shadowed trees than the baseline. Similar trends were observed for UMFusion and MFEIF. Although RGB-thermal early fusion improved the visibility of BG regions, the detection of tree crowns deteriorated. Overall, the detection performance of all three fusion methods was worse than the baseline.

UDA. As shown in Table 1, existing UDA methods require GT annotations to compute task-specific detection loss during training, which guides the adaptation process. To ensure a fair comparison, we selected three UDA methods compatible with the one-stage object detector RetinaNet, including Attention-based UDA [50], SSTN [38], and DA-RetinaNet [41]. We modified these methods by excluding only the supervised detection loss. For Attention-based UDA, their proposed attention module, which dynamically selects local feature regions for adaptation, was trained using DAT alongside the thermal branch of our model. In the case of SSTN, only the ResNet and thermal pre-layer were fine-tuned using contrastive loss as described in [38]. Similarly, for DA-RetinaNet [41] only global DAT was employed to adapt the ResNet and pre-layer of the thermal branch. Among these three methods, DA-RetinaNet demonstrated the best adaptation to the thermal data distribution. However, its performance was still limited as numerous false positive predictions contributed to the overall insufficient performance. The drawback of these methods lies in the absence of task-aware detection loss during adaptation due to the lack of GT annotations. Consequently, the model cannot learn to extract domain-invariant representations that are meaningful for the detection task. Instead, it primarily learns to deceive the domain discriminators irrespective of the downstream task (i.e., detection). Our proposed method overcomes the limitations of the discussed UDA methods by incorporating task-aware FG FPN feature alignment (abbreviated FG FPN FA) to guide the adversarial adaptation process. As a result, our method using DAT and FG FPN FA outperforms the baseline RGB-only detector as well as existing SOTA methods with an AP50 of 54.13% and AR100 of 25.76%, with 19.09% of shadowed trees successfully detected (almost doubling the success rate of the baseline). Particularly, our entirely self-supervised method performs comparably to the supervised fine-tuning method without requiring labor-intensive manual labeling. The feature fusion process in our method selectively enhances features in the BG regions using thermal-extracted features. Importantly, this fusion does not have an adverse effect on FG performance, which distinguishes our method from existing early fusion approaches. Additionally, our multi-modal fusion leverages available data from both domains for detection, unlike single-domain image-to-image translation methods.

4.4. Ablation Study

A systematic ablation analysis of the proposed method is presented in Table 3. It includes five different configurations: (i) FG FPN FA using our classic image masking (CIM) with no DAT applied to the ResNet model, (ii) ResNet DAT with FPN FA and no masking (aligning all regions of feature maps), (iii) ResNet DAT with FG FPN FA and a different masking (using baseline detector predictions as FG and the rest as BG), (iv) Pixel-wise ResNet DAT (discriminators output domain labels for each pixel and consider loss only for FG pixels) with FG FPN FA using CIM, and (v) Proposed ResNet DAT with FG FPN FA using CIM (ShadowSense).

According to the results, the following key inferences can be made: 1) UDA through DAT is crucial (from (i) & (v)): relying solely on FG FPN FA led to inferior performance compared to the baseline, indicating that the thermal branch did not effectively learn to extract domain-invariant representations without DAT. 2) FG masking for FPN alignment is crucial (from (ii), (iii) & (v)): Even with DAT, aligning whole feature maps slightly decreased performance compared to the baseline – aligning features of a tree visible only in the thermal image with BG features from the RGB image interferes with training. 3) Our mask generation method outperforms RGB detector-predicted mask generation (from (iii) & (v)): The proposed masking detects significantly more shadowed trees than this alternate masking, showing superior performance for FPN FA and fusion. 4) FG masking is unnecessary for DAT (from (iv) & (v)): Pixel-wise domain classifiers with loss computation restricted to FG regions resulted in slightly worse performance than global DAT, likely due to the usage of less available data (only FG pixels vs. all pixels) in the same number of training iterations.

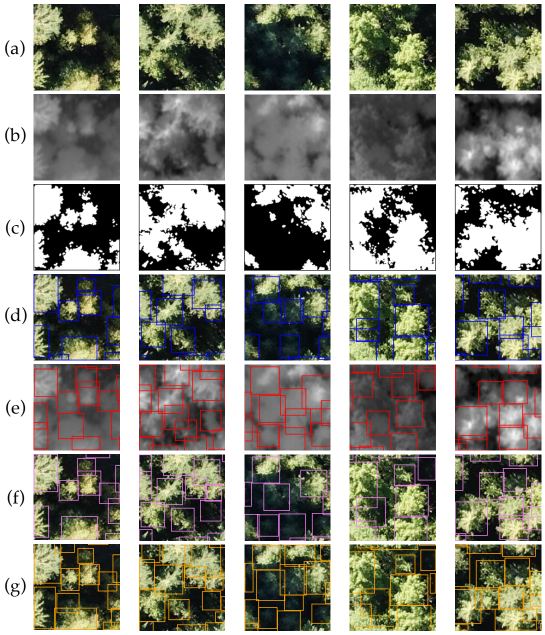


Figure 4. **Detection Results**. Each column shows (a) RGB image, (b) Thermal image, (c) Generated mask; and predictions by (d) Baseline [56], (e) Our DAT-adapted thermal branch, (f) Proposed ShadowSense, and (e) Ground truth. Best viewed in color and zoom-in.

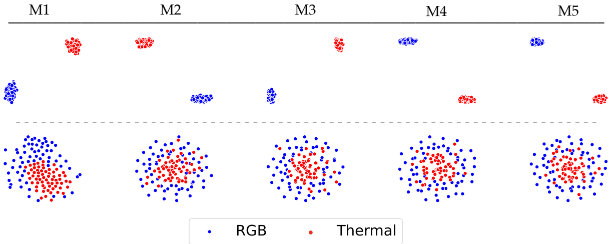


Figure 5. **t-SNE Visualization** of RGB-thermal FPN features: (top row) before training and (bottom row) after training.

4.5. Qualitative Results

Detection Performance. The performance of the proposed method is compared to the off-the-shelf RGB detector [56] in Fig. 4. Two outputs for our method are shown: one from thermal FPN feature maps (isolated thermal branch) and the other from the fused feature maps. The thermal branch detects shadowed trees in the BG that were missed by the baseline, but there is a decline in the FG performance. In the fused output, the BG detections are accurately propagated while maintaining the baseline performance in the FG. Overall, our method outperforms the baseline by comprehensively improving the detection results. Additional qualitative results can be found in the supplementary material.

Feature Space Visualization. Visualizing the FPN feature maps for the testing set in Fig. 5 shows the initial disparity between the RGB-thermal modalities before our training procedure. After training, however, they become indistinguishable as domain-invariant feature maps are aligned and can be directly averaged for fusion during inference.

5. Conclusions

We presented a novel shadow-agnostic ITCD method and a challenging paired RGB-thermal dataset to address the limitations of existing RGB-trained detectors. Our method exploits DAT and FG FPN feature alignment to learn domain-invariant representations and match visible tree crowns in RGB and thermal modalities. Unlike existing adaptation methods, our approach does not require source annotations for task-aware supervision during training, but instead relies on the registered nature of image pairs for aligning features of visible FG regions. Our approach effectively detects small trees hidden in the shadow of neighboring taller trees by fusing complementary thermal information. Further, our dataset comprises aligned RGB-thermal drone image pairs that can stimulate future research in challenging ITCD scenarios. Experimental comparisons demonstrate the superiority of our proposed method over the baseline RGB-trained detector and SOTA image fusion- and UDA-based techniques.

Acknowledgments. We thank fRI Research - MPB Ecology Program for funding; Weyerhauser Company for Cynthia cutblock access; Guillermo Castilla, Mike Gartrell, Jim Weber, and Steven Wagers of NRCan for data acquisition.

References

[1] Ibrahim Batuhan Akkaya, Fazil Altinel, and Ugur Halici. Self-training guided adversarial domain adaptation for thermal imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4322–4331, 2021. 2

[2] Mirela Beloiu, Lucca Heinzmann, Nataliia Rehush, Arthur Gessler, and Verena C. Griess. Individual tree-crown detection and species identification in heterogeneous forests using aerial RGB imagery and deep learning. *Remote Sensing*, 15(5):1463, Mar. 2023. 1

[3] Serge Beucher and Fernand Meyer. The morphological approach to segmentation: The watershed transformation. In Edward Dougherty, editor, *Mathematical Morphology in Image Processing*, pages 433–481. Marcel Dekker Inc., New York, 1993. 4

[4] Junxiao Chen, Jia Wei, and Rui Li. Targan: Target-aware generative adversarial networks for multi-modality medical image translation. *arXiv preprint arXiv:2105.08993*, 2021. 2, 3

[5] Simon Ecke, Jan Dempewolf, Julian Frey, Andreas Schwaller, Ewald Endres, Hans-Joachim Klemmt, Dirk Tiede, and Thomas Seifert. UAV-based forest health monitoring: A systematic review. *Remote Sensing*, 14(13):3205, 2022. 1

[6] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009. 3

[7] Ayana Fujimoto, Chihiro Haga, Takanori Matsui, Takashi Machimura, Kiichiro Hayashi, Satoru Sugita, and Hiroaki Takagi. An end to end process development for uav-sfm based forest monitoring: Individual tree detection, species classification and carbon dynamics simulation. *Forests*, 10(8):680, 2019. 1

[8] Lu Gan, Connor Lee, and Soon-Jo Chung. Unsupervised RGB-to-thermal domain adaptation via multi-domain attention network. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2023. 2

[9] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France, July 2015. PMLR. 4, 6

[10] Peng Gao, Tian Tian, Tianming Zhao, Linfeng Li, Nan Zhang, and Jinwen Tian. GF-detection: Fusion with GAN of infrared and visible images for vehicle detection at nighttime. *Remote Sensing*, 14(12):2771, 2022. 2

[11] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023. 2, 3, 6, 7

[12] Qinling Guo, Wujie Zhou, Jingsheng Lei, and Lu Yu. TSFNet: Two-stage fusion network for RGB-t salient object detection. *IEEE Signal Processing Letters*, 28:1655–1659, 2021. 2, 3

[13] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Y. Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 5

[14] Qihui Han and Cheolkon Jung. Deep selective fusion of visible and near-infrared images using unsupervised u-net. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2022. 2

[15] S. N. H. Syed Hanapi, S. A. A. Shukor, and J. Johari. A review on remote sensing-based method for tree detection and delineation. *IOP Conference Series: Materials Science and Engineering*, 705(1):012024, Nov. 2019. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Sept. 2018. 2, 3

[18] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3496–3504, Oct. 2021. 5

[19] Rudraksh Kapil, Guillermo Castilla, Seyed Mojtaba Marvasti-Zadeh, Devin Goodsman, Nadir Erbilgin, and Nilanjan Ray. Orthomosaicking thermal drone images of forests via simultaneously acquired RGB images. *Remote Sensing*, 15(10):2653, 2023. 5

[20] Rudraksh Kapil, Seyed Mojtaba Marvasti-Zadeh, Devin Goodsman, Nilanjan Ray, and Nadir Erbilgin. Classification of bark beetle-induced forest tree mortality using deep learning. *arXiv preprint arXiv:2207.07241*, 2022. 1

[21] Mehran Khodabandeh, Arash Vahdat, Mani Ranjbar, and William G. Macready. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 480–490, Oct. 2019. 2

[22] My Kieu, Andrew D. Bagdanov, Marco Bertini, and Alberto del Bimbo. Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *Computer Vision – ECCV 2020*, pages 546–562. Springer International Publishing, 2020. 2

[23] Yeong-Hyeon Kim, Ukcheol Shin, Jinsun Park, and In So Kweon. MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. *IEEE Robotics and Automation Letters*, 6(4):6497–6504, 2021. 2

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017. 6

[25] Anqi Li, Dongxu Ye, Erli Lyu, Shuang Song, Max Q.-H. Meng, and Clarence W. de Silva. RGB-thermal fusion network for leakage detection of crude oil transmission pipes. In *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019. 3

[26] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. Explicit attention-enhanced fu-

sion for RGB-thermal perception tasks. *IEEE Robotics and Automation Letters*, 8(7):4060–4067, July 2023. 3

[27] Guibiao Liao, Wei Gao, Ge Li, Junle Wang, and Sam Kwong. Cross-collaborative fusion-encoder network for robust RGB-thermal salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7646–7661, Nov. 2022. 2, 3

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2017. 2, 3, 6

[29] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5802–5811, June 2022. 2, 5, 7

[30] Jinyuan Liu, Xin Fan, Ji Jiang, Risheng Liu, and Zhongxuan Luo. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):105–119, 2022. 2, 7

[31] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *arXiv preprint arXiv:2211.10960*, 2022. 2

[32] Fuya Luo, Yunhan Li, Guang Zeng, Peng Peng, Gang Wang, and Yongjie Li. Thermal infrared image colorization for nighttime driving scenes with top-down guided attention. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15808–15823, 2022. 2, 3, 6, 7

[33] Chengjin Lyu, Patrick Heyer, Bart Goossens, and Wilfried Philips. An unsupervised transfer learning framework for visible-thermal pedestrian detection. *Sensors*, 22(12):4416, 2022. 2

[34] Seyed Mojtaba Marvasti-Zadeh, Devin Goodsman, Nilanjan Ray, and Nadir Erbilgin. Early detection of bark beetle attack using remote sensing and machine learning: A review. *arXiv preprint arXiv:2210.03829*, 2022. 1

[35] Seyed Mojtaba Marvasti-Zadeh, Devin Goodsman, Nilanjan Ray, and Nadir Erbilgin. Crown-CAM: Interpretable visual explanations for tree crown detection in aerial images. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. 2

[36] Giulio Mattolin, Luca Zanella, Elisa Ricci, and Yiming Wang. Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 423–433, Jan. 2023. 2

[37] Fatemeh Moradi, Farzaneh Dadrass Javan, and Farhad Samadzadegan. Potential evaluation of visible-thermal UAV image fusion for individual tree detection based on convolutional neural network. *International Journal of Applied Earth Observation and Geoinformation*, 113:103011, 2022. 1, 2, 3

[38] Farzeen Munir, Shoaib Azam, and Moongu Jeon. SSTN: Self-supervised domain adaptation thermal object detection for autonomous driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021. 2, 4, 7

[39] Masanori Onishi and Takeshi Ise. Explainable identification and mapping of trees using UAV RGB image and deep learning. *Scientific Reports*, 11(1), 2021. 1

[40] Poojan Oza, Vishwanath A. Sindagi, Vibashan Vishnukumar Sharmini, and Vishal M. Patel. Unsupervised domain adaptation of object detectors: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–24, 2023. 2

[41] Giovanni Pasqualino, Antonino Furnari, Giovanni Signorello, and Giovanni Maria Farinella. An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites. *Image and Vision Computing*, 107:104098, 2021. 2, 4, 7

[42] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. *arXiv preprint arXiv:1812.04798*, 2019. 2

[43] Pierre J. Soille and Marc M. Ansoult. Automated basin delineation from digital elevation models using mathematical morphology. *Signal Processing*, 20(2):171–182, 1990. 4

[44] Shaoyue Song, Zhenjiang Miao, Hongkai Yu, Jianwu Fang, Kang Zheng, Cong Ma, and Song Wang. Deep domain adaptation based multi-spectral salient object detection. *IEEE Transactions on Multimedia*, 24:128–140, 2022. 2, 3

[45] Yuxiang Sun, Weixun Zuo, and Ming Liu. RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 3

[46] Richard Szeliski. *Computer Vision Algorithms and Applications 2nd Edition*. Springer London, 2021. 4

[47] Alexander Toet. TNO image fusion dataset. *figshare 10.6084/M9.FIGSHARE.1008029.V2*, 2014. 5, 7

[48] Zhengzheng Tu, Zhun Li, Chenglong Li, Yang Lang, and Jin Tang. Multi-interactive dual-decoder for RGB-thermal salient object detection. *IEEE Transactions on Image Processing*, 30:5678–5691, 2021. 2, 3

[49] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014. 4

[50] Vidit Vidit and Mathieu Salzmann. Attention-based domain adaptation for single-stage detectors. *Machine Vision and Applications*, 33(5), July 2022. 2, 4, 7

[51] Vibashan VS, Domenick Poster, Suya You, Shuowen Hu, and Vishal M. Patel. Meta-uda: Unsupervised domain adaptive thermal object detection using meta-learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1412–1423, Jan. 2022. 2

[52] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022. 2, 7

[53] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow

detection and shadow removal. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1788–1797, 2018. 7

[54] Jie Wang, Kechen Song, Yanqi Bao, Liming Huang, and Yunhui Yan. CGFNet: Cross-guided fusion network for RGB-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2949–2961, 2022. 2, 3

[55] Zhishe Wang, Wenyu Shao, Yanlin Chen, Jiawei Xu, and Xiaoqin Zhang. Infrared and visible image fusion via interactive compensatory attention adversarial learning. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 2, 3

[56] Ben G. Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11):1309, June 2019. 1, 2, 3, 4, 6, 7, 8

[57] Ben G. Weinstein, Sergio Marconi, and Ethan White. Training data for the NEON tree evaluation benchmark. *Zenodo 10.5281/zenodo.5912107*, Jan. 2022. 7

[58] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 2

[59] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. FusionDN: A unified densely connected network for image fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12484–12491, 2020. 5

[60] Yuchen Yang and Nilanjan Ray. Foreground-focused domain adaption for object detection. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021. 2

[61] Qiang Zhang, Tonglin Xiao, Nianchang Huang, Dingwen Zhang, and Jungong Han. Revisiting feature fusion for RGB-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1804–1818, 2021. 2, 3

[62] Xingchen Zhang, Ping Ye, and Gang Xiao. Vifb: A visible and infrared image fusion benchmark. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 468–478, 2020. 5

[63] Haotian Zhao, Justin Morgenroth, Grant Pearse, and Jan Schindler. A systematic review of individual tree crown detection and delineation with convolutional neural networks (CNN). *Current Forestry Reports*, 2023. 1, 2, 7

[64] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13955–13965, Mar. 2023. 2, 7

[65] Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *Computer Vision – ECCV 2020*, pages 54–69. Springer International Publishing, 2020. 2

[66] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13766–13775, June 2020. 2

[67] Wujie Zhou, Qinling Guo, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. ECFFNet: Effective and consistent feature fusion network for RGB-t salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1224–1235, Mar. 2022. 2, 3